CS7IS3: Assignment 1 Crandfiled search engine

Mingwei Shi
Trinity college dublin
mshi@tcd.ie

I. Introdcution

Parser-¿Index-¿Search -¿Evaluation

II. Implementation

A. Parser

Finite state machine to process

B. Indexing

Feature type index:

Only two feature are indexed by Textfild that could be tokeinizedion including title and description. Since the rest of feature shold not be separate

Customiser :tokeinze+lowecaset+poesses remvoal+ large stop words removal.

C. Searcher

Simairty use: VSM;BM25;LMDIRCHATE

III. Evaltion

In this section, I would like to identify the discriminating features combination that has an impact on mean precision value, what is the influential analyzer that has an impact on mean precision value, and what is the influential similarity matching algorithm that has impact on the mean precision value.

a1 for analyzer : Standard
a2 for analyzer : Customized
s1 for Simlairty : Vector space
s2 for Simlairty : BM25
s3 for Simlairty : LMDirchite
f2 for Descirption feature only(.W) : Vector space
f3 for combbing with title+ desritpikn(.T and .W) feature only(.W) :

Obersvation for analyzer: When add more stopword,the recall decrease as the key term are smller;

Obersvation for features:The content of descprtion(.W) is so informative that introducing extra content such as title(.T) would not add the precision.As the information entrpoy might not increase.

Obersvation for simalirty: BM25 is the best

TABLE I
Comparison between different analyzers, similarity, and feature.

Acronym Name	MAP
a1-s1-f2	0,2831
a1-s1-f3	0,2831
a1-s2-f2	0,2945
a1-s2-f3	0,2945
a1-s3-f2	0,231
a1-s3-f3	0,231
a2-s1-f2	0,1667
a2-s1-f3	0,1667
a2-s2-f2	0,1616
a2-s2-f3	0,1616
a2-s3-f2	0,1121
a2-s3-f3	0,1121

References